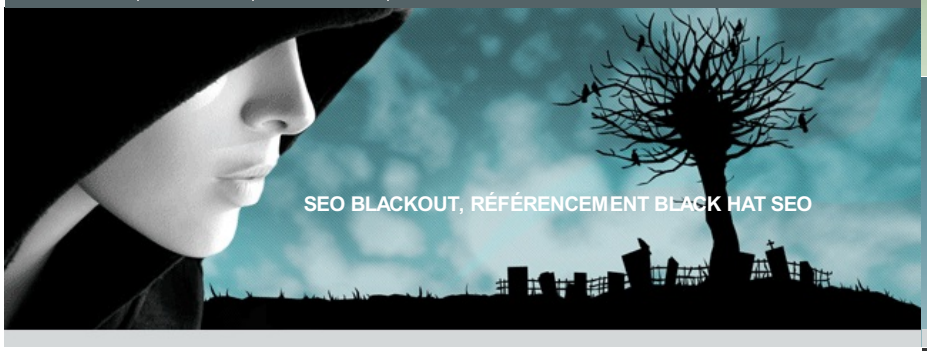




SEO HIGH LEVEL Formation SEO : Méthodes avancées 24 - 25 - 26 Mars 2014

# Extraire les urls des résultats de Google

Ecrit par **admin** le Dimanche 26 octobre 2008 à 15:12 - Catégorie **Scraping**,**Scripts PHP**

**Comment extraire / parser les urls de résultats de recherche de Google ?**

Cela fait plusieurs fois qu'on me demande comment extraire les urls de résultats de recherche de Google dans le but de les utiliser pour diverses raisons : recherche de pages de soumission d'annuaires, recherche des pages de blog avec commentaires, pages de forums, commande link ou site…

Voici donc un script php qui permet d'extraires ces urls de Google.

Différentes options sont proposées, comme le nombre de pages à extraire ou le moteur Google où se fera l'extraction ainsi que la langue du moteur.

**Code source du script :**

Voici le code du script php à copier/coller dans un fichier et à uploader sur votre serveur ou à utiliser en local :

```
1  <?php
2  //Script php : Extraction urls Résultats Google
3  // Copyrights 2008 Seoblackout.com
4  //http://www.seoblackout.com
5  //Url script : http://www.seoblackout.com/2008/10/26/extraire-resultats-google
6
7  @set_time_limit(0);
8  $useragent='Mozilla/5.0';
9  $regex='<h3 class="r">';
10 //extension et langue google
11 if ((isset($_POST['ext'])) && ($_POST['ext']!=''))
12 {
13  $ext=strip_tags($_POST['ext']);
14  if ($ext=='fr') {
15  $lang='fr';
16  }
17  else if ($ext=='com') {
18  $lang='en';
19  }
20  else if ($ext=='es') {
21  $lang='es';
22  }
23  else if ($ext=='de') {
24  $lang='de';
25  }
26  else if ($ext=='ca') {
27  $lang='en';
28  }
29  else if ($ext=='ca2') {
30  $ext='ca';
31  $lang='fr';
32  }
33  else {
34  $ext='fr';
35  $lang='fr';
36  }
37 }
38 else
39 {
40  // par défaut, on recherche sur google.fr et lang=fr
41  $ext='fr';
42  $lang='fr';
43 }
44 //nombre de page à extraire :
45 if ((is_numeric($_POST['pages'])) && ($_POST['pages']!=''))
46 {
47  $c=(strip_tags($_POST['pages'])-1);
48 }
49 else
```



Copyrights SeoBlackOut 2007 - 2014 ©
Propulsé par WordPress
28 requêtes bdd.
Page générée en 0,824 secondes.

```
50  {
51   //10 pages par défaut, on part de 0 donc 10 pages = 9
52   $c=9;
53  }
54
55  if ((isset($_POST['kw'])) && ($_POST['kw']!=''))
56  {
57   $kw=trim(strip_tags($_POST['kw']));
58   $pagenum = 0;
59   $googlefrurl = "http://www.google.".$ext."/search?hl=".$lang."&q=" . urlencod
60   $url_new = '';
61
62   while($pagenum <= $c)
63   {
64     if (function_exists('curl_init'))
65     {
66      $ch = curl_init();
67      curl_setopt($ch, CURLOPT_USERAGENT, $GLOBALS['useragent']);
68      curl_setopt($ch, CURLOPT_URL,$googlefrurl);
69      curl_setopt($ch, CURLOPT_RETURNTRANSFER,1);
70      $result=curl_exec ($ch);
71      curl_close ($ch);
72     }
73     else
74     {
75      $result= file_get_contents($googlefrurl);
76     }
77    preg_match_all('/'.$regex.'<a href="(.*?)"/si', $result, $matches);
78     $i = 0;
79     $n = count($matches[1]);
80     $pagenum++;
81     $pagenum2 = $pagenum.'0';
82     $googlefrurl = "http://www.google.".$ext."/search?hl=".$lang."&q=" . urlenc
83
84     while($i <= $n)
85     {
86      $url_new1 = addslashes($matches[1][$i]);
87      $url_new .= trim($matches[1][$i])."<br>";
88      $i++;
89      flush();
90     }
91   }
92   $url_new .= '';
93  }
94  ?>
95  <!DOCTYPE html PUBLIC "-//W3C//DTD XHTML 1.0 Transitional//EN" "http://www.w3.
96  <html xmlns="http://www.w3.org/1999/xhtml">
97  <head>
98  <meta http-equiv="Content-Type" content="text/html; charset=iso-8859-1" />
99  <title>Extraction Urls résultats Google</title>
100 </head>
101 <body>
102 <h1>Extraction Urls résultats Google</h1>
103 <div>
104 <?php
105 if ((isset($_POST['kw'])) && ($_POST['kw']!='')) {
106  echo "<h3>Keyword : ",strip_tags($kw),"</h3>";
107  echo "<h6>Google Results</h6>";
108  echo str_replace('<br><br>','<br>',$url_new);
109 }
110 ?>
111 <form method="POST" action="<?php echo strip_tags($_SERVER['REQUEST_URI']) ;?>
112 <p>Saisir vos mots-clés :</p>
113 <input name="kw" type="text" size="100" value="<?php if (isset($_POST['kw']))
114 <p>Nombre de pages à extraire (optionel => 10 pages par défaut):</p>
115 <input name="pages" type="text" size="3" maxlength="3" value="<?php if (isset(
116 <p>Moteur :</p>
117 <select name="ext">
118 <option value="fr"<?php if (($_POST['ext'])=='fr') {echo ' selected';} ?>>Goog
119 <option value="com"<?php if (($_POST['ext'])=='com') {echo ' selected';} ?>>Go
120 <option value="es"<?php if (($_POST['ext'])=='es') {echo ' selected';} ?>>Goog
121 <option value="de"<?php if (($_POST['ext'])=='de') {echo ' selected';} ?>>Goog
122 <option value="ca"<?php if (($_POST['ext'])=='ca') {echo ' selected';} ?>>Goog
123 <option value="ca2"<?php if (($_POST['ext'])=='ca2') {echo ' selected';} ?>>Go
124 </select>
125 <p><input type="submit" value="Go" name="go">
126 <input type='button' value='Annuler' onclick='location.href="<?php echo strip_
127 </form>
128 <br />
129 <small>Script proposé par SEO BlackOut : <a href="http://www.seoblackout.com/"
130 </div>
131 </body>
132 </html>
```

**Démo :**

*Dans la démo, le nombre de résultats extraits est limité à 3 pages uniquement !*

http://www.seoblackout.com/scripts/extraire-urls-google.php

Tweet

**Articles liés :**

- Extraire les urls des résultats de Bing

- Google propose d'afficher les résultats de recherche en anglais pour certaines requêtes
- Référencement Google, nouvel ajout dans les résultats ?
- Extraire les backlinks référencés par Alexa
- Google Wonder Wheel PHP Script
- Identifier les pages en résultats complémentaires

Commentaires (29)
Catégorie: Scraping,Scripts PHP
Besoin d'une prestation en référencement ? Nous contacter

## 29 Commentaires

Commentaire par **Mode enfant Magazine**
Dimanche, 26 octobre , 2008 à 23:38

Bon script
Attention cependant au captcha pour requêtes automatisées

Commentaire par **admin**
Lundi, 27 octobre , 2008 à 0:06

Yep à utiliser en local si ip variable ou faire tourner sur différents serveurs ou passer par proxy ou ne pas abuser… ou connaitre les techniques secrètes

Commentaire par **tictact**
Lundi, 27 octobre , 2008 à 12:52

très bon script
thanks paulo

Commentaire par **guizme555**
Lundi, 27 octobre , 2008 à 17:46

Merki encore une fois pour ce script Je fais tourner !

Commentaire par **Rudy**
Mardi, 28 octobre , 2008 à 22:14

Yo

En voyant le titre je me suis dit cool voila du contenu :p
Mais en testant je me suis rendu compte que le script recupère les url uniquement,a quoi ca sert ?

J'ai cru qu'avec on pouvait choper le contenu des 1ers pages de résultats

Commentaire par **Louis-Philippe Huberdeau**
Mardi, 28 octobre , 2008 à 22:25

Tu peux te sauver beaucoup d'effort en utilisant les API…

Commentaire par **admin**
Mercredi, 29 octobre , 2008 à 2:10

@Louis-Philippe Huberdeau : l'api est intéressante, mais pas pour l'usage souhaité ici, api limitée…

@Rudy, ça sert à te faire de belles listes d'urls à spammer automatiquement par exemple.

Commentaire par **lo**
Mercredi, 29 octobre , 2008 à 9:28

Parfait. Merci pour ce script !

Commentaire par **Karedas**
Mercredi, 29 octobre , 2008 à 13:37

*J'ai cru qu'avec on pouvait choper le contenu des 1ers pages de résultats*
Une fois que tu as les urls stockés, facile de faire un script spider qui va les explorer et en extraire le contenu.
Il vaut mieux éviter extraction google et crawl en meme temps sans quoi t'en a vite pour plusieurs minutes de travail par recherche.

Je suis très, mais alors très très, surpris que SEO BLACKOUT ne sache pas que, pour éviter le ban de l'ip par google, il suffit de remplacer 'search' par 'custom'.

Commentaire par **admin**
Jeudi, 30 octobre , 2008 à 0:11

T'inquiètes Karedas, on utilise custom aussi qui au passage n'est pas suffisant, il ne suffit pas de

remplacer custom par search, il y a d'autres paramètres à passer pour ne pas se faire bloquer mais ça on le garde en interne, ou alors je laisse le développeur du script le présenter sur SEO Blackout s'il le souhaite (mais ça m'étonnerait qu'il accepte…)

Commentaire par **BlackMelvyn**
Vendredi, 31 octobre , 2008 à 16:35

En effet, même s'il y a un intérêt pédagogique très important, les meilleures astuces Black Hat sont aussi les moins connues
Peut-être que nous présenterons le script plus tard car la méthode est intéressante, mais pour le moment, c'est un peu trop précieux à mes yeux pour le dévoiler au public… et n'oublions pas qu'ici c'est aussi **Don't be evil**
Sorry
Je pense cependant que Paul en a méchamment aidé ici avec ce script y'a plus qu'à

Commentaire par **Jice**
Mercredi, 4 février , 2009 à 19:56

Je viens de me mettre au python, et j'ai voulu commencer par l'adaptation de ce petit script. Le résultat est à voir ici :
http://jice.lavocat.name/blog/2009/02/extraire-urls-des-resultats-google-en-python/

C'est pas un beau rendu dans une page web, mais ça m'a suffit pour apprendre les bases (récupération url, traitement avec Regexp…).
Le script a été un peu modifié pour contourner des problème que j'ai pas résolus autrement. Mon code est donc 'sale'. Si vous avez des commentaires enrichissants à faire je suis preneur. De même qu'un bon plugin de Geshi pour WordPress…

Commentaire par **admin**
Jeudi, 5 février , 2009 à 22:14

Salut Jice,

Je peux pas dire grand chose sur ton script vu que je connais pas python
Par contre merci pour le lien

Pour le plugin wordpress, je peux te proposer WP-Syntax (c'est celui que j'utilise ici) :
http://wordpress.org/extend/plugins/wp-syntax/

Commentaire par **Tigrou**
Mercredi, 17 juin , 2009 à 9:09

Bonjour,

Je me demandais si l'api google search permet d'éviter le ban de l'ip par google?

Merci pour vos lumières

Commentaire par **Olivier**
Dimanche, 26 juillet , 2009 à 9:03

Pour ma part, j'ai créé un script pour connaître la position de son site sur Google pour un mot clé donné :
http://www.olivier-tassel.fr/outils/script-position-site-google-mot-cle

Commentaire par **Jack**
Mardi, 8 septembre , 2009 à 21:22

De la balistique.

- Exactement ce que je cherchais ; je cherche à faire un petit prototype de la même veine en Perl, suis penché sur le pb depuis plusieurs jours, mais ne vois toujours pas comment utiliser cURL ou libcurl pour extraire des URL de retours de requête Google ; en effet le source des pages de requête Google ne fournipas d'URL, mais plutôt des clés internes (à Google) pour chaque URL.

Cf par exemple, sur un shell :
curl 'http://www.google.fr/search?source=ig&hl=fr&rlz=&=&q=ma+requete&btnG=Recherche+Google&meta=lr%3Dlang_fr&aq=f&oq='

- Louis Philippe : il est vrai que des tas d'interfaces, dans tous les langages, sont à disposition sur code.Google pour interroger la « Google Base API ». Les résultats sont donnés sous formes de flux XML, dans un format XML propre à Google : on peut alors exploiter les retours de requête Google de manière extrêmement fine, mais ça m'a l'air fastidieux. Utiliser cURL comme ici, en manipulant directement la syntaxe (http) des pages de resultat Google, me semble au contraire infiniement plus simple.

@+ Bravo pour ton code, qui a un intérrêt énorme par rapport à ce qu'on trouve d'autre sur le sujet : il marche très bien !!!

Commentaire par **Stéphane**
Lundi, 22 février , 2010 à 8:56

Salut,

J'aurais aimer savoir quelque chose avant tout, est ce que google autorise ce genre technique à parser leur résultat ?

Car au final ou pourrait presque en allant loin des les regex ce faire son propre moteur. Et puis ont utilise leur requêtes.

Risque ton quelque chose si il repère cela ? Risque ton d'être backlisté ou quelque chose du genre ?

J'ai éssayé celui de Bing, ma foi bien et facile à mettre en place, sauf le nombre de résultat par page.

Enfin merci de votre réponse.

Bye

Commentaire par **Expert**
Mercredi, 7 avril , 2010 à 10:34

Bonjour,

Pour ma part, je l'ai mis en place sur un serveur FTP.
Mais lorsque je lance une requête, il ne m'affiche rien sous « Google Results ».
cela marche-t-il toujours pour vous ?

Commentaire par **nicolas laustriat**
Vendredi, 7 mai , 2010 à 13:05

Ah! j'utilise cet extract depuis sa publication, il a toujours bien focntionné. Mais depuis hier nada…est ce que cela aurait un rapport avec la sortie « officielle » de Jazz?

Commentaire par **oups**
Mercredi, 12 mai , 2010 à 16:31

Le script ne semble plus fonctionner depuis la nouvelle interface de Google.
Dommage le script était assez efficace et simple à mettre en place…

Commentaire par **Emma**
Lundi, 17 mai , 2010 à 18:14

Re-Bonjour,

Désolée pour le « spam », je pense avoir compris pourquoi je n'arrive pas à utiliser vos scripts, cela doit être à cause de la nouvelle version de Google.
Pouvez vous m'indiquez ce qui est à modifier dans chaque script (surement la même chose, ceux-ci utilisant tous le scrap sur google).

Merci

Commentaire par **victor**
Mardi, 18 mai , 2010 à 11:51

Ca ne fonctionne plus
google a changé quelque chose !

a voir svp !

Commentaire par **mac97000**
Samedi, 22 mai , 2010 à 15:00

Euh …

sauf erreur de ma part il suffit de remplacer :

```
$regex='<h3 class=r>';
```

par

```
$regex='<h3 class="r">';
```

à la ligne 9 pour revoir apparaitre la liste des liens …

Commentaire par **nicolas laustriat**
Mardi, 25 mai , 2010 à 10:58

Yes ca marche merci MAC97000

Commentaire par **mac97000**
Mardi, 25 mai , 2010 à 11:35

pas de quoi, en même temps j'ai pas fait grand chose puisque je testais le script

Commentaire par **Remguy**
Vendredi, 23 juillet , 2010 à 17:28

Personnellement j'ai pas réussi à le faire tourner…

Commentaire par **laustriat**
Mardi, 28 septembre , 2010 à 22:26

yep définitivemant c'est un script que j'utilise régulièrement il est top mais je sais pas si vous avez vu ca

aussi (du jour via @rudy su twitter)? http://blog.authoritylabs.com/data-collection-with-google-spreadsheets/

j'ai comparé les résultats (on ne siat jamais) c'est conforme…

Commentaire par **Daniel Roch**
Jeudi, 24 novembre , 2011 à 16:26

Le script ne fonctionne plus ?

Sur mon serveur, il ne me renvoie plus aucune donnée…

Commentaire par **Jocalaz Allan**
Mardi, 24 janvier , 2012 à 12:27

Bonjour,

Est-il possible de récupérer le title affiché par Google. En effet je cherche à travailler un site client qui a 200 URLS indéxées, grâce à ce script je récupère toutes les urls, mais j'aimerais récupérer les titles pour un travail dessus ..

Les commentaires sont fermés pour cet article.